Published in "Digital Signal Processing in Multimedia Systems, Ed. K. K. Parhi and T. Nishitani, Marcell Dekker Inc., 461-485, (1999)

Chapter 18

A review of watermarking principles and practices

Matt L. Miller	Ingemar J. Cox	Jean-Paul M.G. Linnartz Ton Kalker
Signafy Inc. 4 Independence Way Princeton, NJ 08540 USA	NEC Research Institute 4 Independence Way Princeton, NJ 08540 USA	Philips Research Prof. Holstlaan 4 5656 AA Eindhoven The Netherlands

Abstract*

A digital watermark embeds an imperceptible signal into data such as audio, video and images, for a variety of purposes, including captioning and copyright control. As watermarking is increasingly used for a wide variety

*Portions of this paper appeared in the Proceedings of SPIE, Human Vision & Electronic Imaging II, V 3016, pp 92-99, February 1997. Portions are reprinted, with permission, from "Public watermarks and resistance to tampering", I.J. Cox and J.-P. Linnartz, IEEE International Conference on Image Processing, CD-ROM Proc. ©1997 IEEE and from "Some General Methods for Tampering with Watermarks", I. J. Cox and J.-P. Linnartz, IEEE Trans. on Selected Areas of Communications, 16, 4, 587-593, ©1998 IEEE.

of applications, various properties of watermarks, such as how they respond to common signal transformations or deliberate attack, have become important considerations. In this paper, we discuss the important properties of watermarks, and review many approaches.

18.1 Introduction

Digital representations of copyrighted material such as movies, songs, and photographs offer many advantages. However, the fact that an unlimited number of perfect copies can be illegally produced is a serious threat to the rights of content owners. Until recently, the primary tool available to help protect content owners' rights has been encryption. Encryption protects content during the transmission of the data from the sender to receiver. However, after receipt and subsequent decryption, the data is no longer protected and is in the clear.

Watermarking compliments encryption. A digital watermark is a piece of information that is hidden directly in media content, in such a way that it is imperceptible to a human observer, but easily detected by a computer. The principal advantage of this is that the content is inseparable from the watermark. This makes watermarks suitable for several applications, including:

Signatures The watermark identifies the owner of the content. This information can be used by a potential user to obtain legal rights to copy or publish the content from the contact owner. In the future, it might also be used to help settle ownership disputes [†].

Fingerprinting Watermarks can also be used to identify the content buyers. This may potentially assist in tracing the source of illegal copies. This idea has been implemented in the DIVX digital video disk players, each of which places a watermark that uniquely identifies the player in every movie that is played.

Broadcast and publication monitoring As in signaturing, the watermark identifies the owner of the content, but here it is detected by automated systems that monitor television and radio broadcasts, com-

[†]In a recent paper [1] it was shown that the use of watermarks for the establishemnt of ownership can be problematic. It was shown that for a large class of watermarking schemes a so called "counterfeit original" attack can be used to confuse ownership establishment. A technical way out may be the use of *one-way watermark* functions, but the mathematical modelling of this approach is still in its infancy. In practical terms the combined use of a copyright office (along the guidelines of WIPO) and a watermark label might provide sufficiently secure fingerprints.

puter networks, and any other distribution channels to keep track of when and where the content appears. This is desired by content owners who wish to ensure that their material is not being illegally distributed, or who wish to determine royalty payments. It is also desired by advertisers who wish to ensure that their commercials are being broadcast at the times and locations they have purchased.

Several commercial systems already exist which make use of this technology. The MusiCode system provides broadcast monitoring of audio, VEIL-II and MediaTrax provide broadcast monitoring of video. Also, in 1997 a European project by the name of VIVA was started to develop watermark technology for broadcast monitioring.

Authentication Here, the watermark encodes information required to determine that the content is authentic. It must be designed in such a way that any alteration of the content either destroys the watermark, or creates a mismatch between the content and the watermark that can be easily detected. If the watermark is present, and properly matches the content, the user of the content can be assured that it has not been altered since the watermark was inserted. This type of watermark is sometimes referred to as a *vapormark*.

Copy control The watermark contains information about the rules of usage and copying which the content owner wishes to enforce. These will generally be simple rules such as "this content may not be copied", or "this content may be copied, but no subsequent copies may be made of that copy". Devices which are capable of copying this content can then be required by law or patent license to test for and abide by these watermarks. Furthermore, devices that can play the content might test for the watermarks and compare them with other clues, such as whether the content is on a recordable storage device, to identify illegal copies and refuse to play them. This is the application that is currently envisaged for digital video disks (DVD).

Secret communication The embedded signal is used to transmit secret information from one person (or computer) to another, without anyone along the way knowing that this information is being sent. This is the classical application of steganography – the hiding of one piece of information within another. There are many interesting examples of this practice from history, e.g., [2]. In fact, Simmons' work [3] was motivated by the Strategic Arms Reduction Treaty verification. Electronic detectors were allowed to transmit the status (loaded or unloaded) of a nuclear missile silo, but not the position of that silo. It appeared that digital signature schemes which were intended to

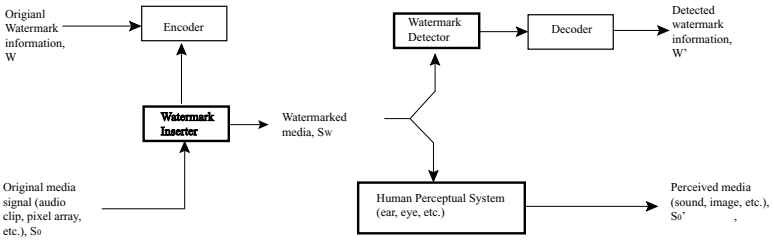

Figure 18.1: Watermarking framework.

verify the integrity of such status message, could be misused as a "subliminal channel" to pass long espionage information.

There are several public-domain and shareware programs available that employ watermarking for secret communication. Rivest [4] has suggested that the availability of this technology casts serious doubt on the effectiveness of government restrictions on encryption, since these restrictions cannot apply to steganography.

These are some of the major applications for which watermarks are currently being considered or used, but several others are likely to appear as the full implications of this technology are realized.

In the next section, we present the basic principles of watermarking and then, in Section 18.3 go on to discuss several properties of watermarking technologies. In section 18.4 we describe a simple watermarking method that then allows for a detailed discussions of robustness (Section18.5) and tamper-resistance (Section18.6). Finally, Section 18.7 gives a brief overview of several watermarking methods.

18.2 Framework

Figure 18.1 shows the basic principle behind watermarking. Watermarking is viewed as a process of combining two pieces of information in such a way that they can be independently detected by two very different detection processes. One piece of information is the media data S_0, such as music, a photograph, or a movie, which will be viewed (detected) by a human observer. The other piece of information is a watermark, comprising an arbitrary sequence of bits, which will be detected by a specially designed watermark detector.

The first step is to encode the watermark bits into a form that will be easily combined with the media data. For example, when watermarking

images, watermarks are often encoded as two- dimensional, spatial patterns. The watermark inserter then combines the encoded representation of the watermark with the media data. If the watermark insertion process is designed correctly, the result is media that appears identical to the original when perceived by a human, but which yields the encoded watermark information when processed by a watermark detector.

Watermarking is possible because human perceptual processes discard significant amounts of data when processing media. This redundancy is, of course, central to the field of lossy compression [5]. Watermarking exploits the redundancy by hiding encoded watermarks in them. A simple example of a watermarking method will illustrate how this can be done. It is well known that changes to the least significant bit of an 8-bit grey-scale image cannot be perceived. Turner [6] proposed hiding a watermark in images by simply replacing the least-significant bit with a binary watermark pattern. The detector looks at only the least- significant bit of each pixel, ignoring the other 7 bits. The human visual system looks at only the 7 most-significant bits, ignoring the least-significant. Thus, the two pieces of information are both perfectly detected from the same data stream, without interfering with one another. The least-significant-bit method of watermarking is simple and effective, but lacks some properties that may be essential for certain applications.

Most watermark detection processes require certain information to insert and extract watermarks. This information can be referred to as a "key" with much the same meaning as is used in cryptography. The level of availability of the key in turn determines who is able to read the watermark. In some applications, it is essential that the keys be widely known. For example, in the context of copy protection for digital video disks (DVD) it is envisaged that detectors will be present in all DVD players and will need to read watermarks placed in all copyrighted video content. In other applications, knowledge of the keys can be more tightly restricted.

In the past, we have referred to these two classes of watermarks as public and private watermarking. However, this could be misleading, given the well known meaning of the term "public" in cryptography. A public-key encryption algorithm involves two secrets; encrypting a message requires knowing one secret, and decrypting a message requires knowing the second. By analogy, a "public watermarking" method should also involve two secrets: inserting a watermark would require knowing one, and extracting would require knowing the second. While watermark messages might be encrypted by a public-key encryption technique before being inserted into media (see, for example, [2]), we know of no watermarking algorithm in which the ability to *extract* a watermark (encrypted or not) requires different knowledge than is required for insertion. In practice, all watermarking

algorithms are more analogous to symmetric cryptographic processes in that they employ only one key. They vary only on the level of access to that key. Thus, in this chapter, we refer to the two classes as "restricted-key" and "unrestricted-key" watermarks.

It should be noted that the framework illustrated in Figure 18.1 is different from the common conceptualization of watermarking as a process of arithmetically adding patterns to media data [7, 8, 9]. When the linear, additive view is employed for public watermarking, the detector is usually conceived of as a signal detector, detecting the watermark pattern in the presence of noise - that "noise" being the original media data. However, viewing the media data as noise does not allow us to consider two important facts: 1) unlike real noise, which is unpredictable, the media data is completely known at the time of insertion, and 2) unlike real noise, which has no commercial value and should be reduced to a minimum, the media data must be preserved. Consideration of these two facts allows the design of more sophisticated inserters.

18.3 Properties of watermarks

There are a number of important characteristics that a watermark can exhibit. These include that the watermark is difficult to notice, survives common distortions, resists malicious attacks, carries many bits of information, can coexist with other watermarks, and requires little computation to insert or detect. The relative importance of these characteristics depends on the application. The characteristics are discussed in more detail next.

Fidelity The watermark should not be noticeable to the viewer nor should the watermark degrade the quality of the content. In earlier work [7, 8], we had used the term "imperceptible", and this is certainly the ideal. However, if a signal is truly imperceptible, then perceptually-based lossy compression algorithms either introduce further modifications that jointly exceed the visibility threshold or remove such a signal.

The objective of a lossy compression algorithm is to reduce the representation of data to a minimal stream of bits. This implies that changing any bit of well encoded data should result in a perceptible difference; otherwise, that bit is redundant. But, if a watermark is to be detectible after the data is compressed and decompressed, the compressed unwatermarked data must be different than the compressed watermarked data, and this implies that the two versions of the data will be perceptibly different once they are decompressed and

viewed. Thus, as compression technology improves, watermarks that survive compression will cause increasingly perceptible differences in data that has been compressed and decompressed.

Early work on watermarking focused almost exclusively on designing watermarks that were imperceptible and therefore often placed watermark signals in perceptually insignificant regions of the content, such as high frequencies or low-order bits. However, other techniques, such as spread spectrum, can be used to add imperceptible or unnoticeable watermarks in perceptually significant regions. And, as is pointed out below, placing watermarks in perceptually significant regions can be advantagous for robustness against signal processing.

Robustness Music, images and video signals may undergo many types of distortions. Lossy compression has already been mentioned, but many other signal transformations are also common. For example, an image might be contrast enhanced and colors might be altered somewhat, or an audio signal might have its bass frequencies amplified. In general, a watermark must be robust to transformations that include common signal distortions as well as digital-to-analog and analog-to-digital conversion and lossy compression. Moreover, for images and video, it is important that the watermark survive geometric distortions such as translation, scaling and cropping.

Note that robustness actually comprises two separate issues: 1) whether or not the watermark is still present in the data after distortion and 2) whether the watermark detector can detect it. For example, watermarks inserted into images by many algorithms remain in the signal after geometric distortions such as scaling, but the corresponding detection algorithms can only detect the watermark if the distortion is first removed. In this case, if the distortion cannot be determined and/or inverted, the detector cannot detect the watermark even though the watermark is still present albeit in a distorted form.

Figure 18.2 illustrates one way of conceptualizing robustness. Here we imagine all the possible signals (images, audio clips, etc.) arranged in a two-dimensional space. The point S_0 represents a signal without a watermark. The point S_w represents the same signal with a watermark. The dark line shows the range of signals that would all be detected as containing the same watermark as S_w, while the dotted line indicates the range of distorted versions of S_w that are likely to occur with normal processing. This dotted line is best thought of as a contour in a probability distribution over the range of possible distortions of S_w. If the overlap between the watermark detection region

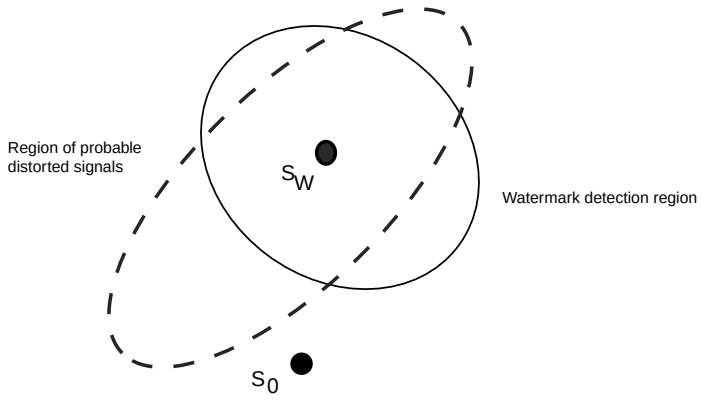

Figure 18.2: Watermark robustness

and the range of likely distorted data is large, then the watermark will be robust.

Of course, in reality, it would be impossible to arrange the possible signals into a two-dimensional space in which the regions outlined in Figure 18.2 would be contiguous, but the basic way of visualizing the robustness issue applies to higher dimensional spaces as well.

A more serious problem with Figure 18.2 is that it is very difficult to determine the range of likely distortions of S_w, and, therefore, difficult to use this visualization as an analytical guide in designing watermarking algorithms. Rather than trying to predetermine the distribution of probable distorted signals, Cox *et al* [7, 8] have argued that robustness can be attained if the watermark is placed in perceptually significant regions of signals. This is because, when a signal is distorted, its fidelity is only preserved if its perceptually significant regions remain intact, while perceptually insignificant regions might be drastically changed with little effect on fidelity. Since we care most about the watermark being detectible when the media signal is a reasonable match with the original, we can assume that distortions which maintain the perceptually significant regions of a signal are likely, and represent the range of distortions outlined by the dotted line in Figure 18.2. Section 18.5 details particular signal processing operations and their effects on detector performance.

Fragility In some applications, we want exactly the opposite of robustness.

Consider, for example, the use of physical watermarks in bank notes. The point of these watermarks is that they do not survive any kind of copying, and therefore can be used to indicate the bill's authenticity. We call this property of watermarks, fragility. Offhand, it would seem that designing fragile watermarking methods is easier than designing robust ones. This is true when our application calls for a watermark that is destroyed by every method of copying short of perfect digital copies (which can never effect watermarks). However, in some applications, the watermark is required to survive certain transformations and be destroyed by others. For example, a watermark placed on a legal text document should survive any copying that doesn't change the text, but be destroyed if so much as one punctuation mark of the text is moved.

This requirement is not met by digital signatures developed in cryptology, which verify bit-exact integrity but cannot distinguish between various degrees of acceptable modifications.

Tamper-resistance Watermarks are often required to be resistant to signal processing that is solely intended to remove them, in addition to being robust against the signal distortions that occur in normal processing. We refer to this property as tamper-resistance. It is desirable to develop an analytical statement about watermark tamper-resistance. However, this is extremely difficult, even more so than in cryptography, because of our limited understanding of human perception. A successful attack on a watermark must remove the watermark from a signal without changing the perceptual quality of the signal. If we had perfect knowledge of how the relevant perceptual process behaved and such models would have tractable computation complexity, we could make precise statements about the computational complexity of tampering with watermarks. However, our present understanding of perception is imperfect, so such precise statements about tamper-resistance cannot yet be made.

We can visualize tamper-resistance in the same way that we visualize robustness, see Figure 18.3. Here, the dotted line illustrates the range of signals that are perceptually equivalent to S_0. As in Figure 18.2, this dotted line should be thought of as a contour in a probability distribution, this time the probability that a signal will be perceived as equivalent to S_0 by a randomly chosen observer. In theory, an attacker who precisely knows the range of this dotted line, as well as the range of the black line (the watermark detection region), could choose a new signal which would be perceptually equivalent to S_0 but would not contain the watermark. The critical issue here is how well known are

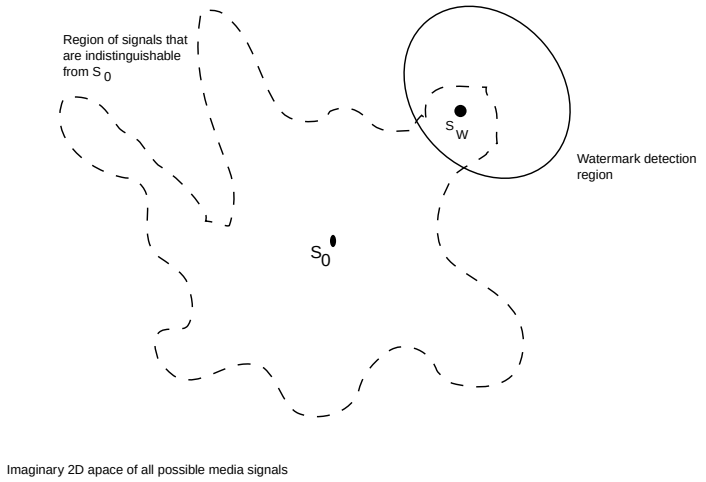

Figure 18.3: Tamper resistance

these two regions. We can assume an attacker does not have access to S_0 (otherwise, she/he would not need to tamper with S_w) so, even if a perfect perceptual model was available, the tamperer could not have perfect knowledge of the region of perceptually equivalent signals. However, the range of signals which are perceptually equivalent to S_0 has a large overlap with those that are perceptually equivalent to S_w, so, if an attacker finds an unwatermarked signal perceptually equivalent to S_w, it is likely to be equivalent to S_0 as well. The success of this strategy depends on how close S_w is to the dotted line. Tamper resistance will be elaborated upon in Section 18.6.

Key restrictions An important distinguishing characteristic is the level of restriction placed on the ability to read a watermark. As explained in earlier sections, we describe watermarks in which the key is available to a very large number of detectors as "unrestricted-key" watermarks, and those in which keys are kept secret by one or a small number of detectors as "restricted-key" watermarks.

While the difference between unrestricted-key and restricted-key is primarily a difference in usage, algorithms differ in their suitability for these two usages. For example, some watermarking methods (e.g. [10]) create a unique key for each piece of data that is watermarked. Such algorithms can be used for restricted-key applications, where the owner of the original data can afford to keep a database of keys for

all the data that has been watermarked. But they cannot be used for unrestricted-key applications, since this would require every detector in the world having a complete list of all the keys. Thus, algorithms for use as unrestricted-key systems must employ the same key for every piece of data.

An unrestricted-key algorithm also must be made resistant to a wider variety of tampering attacks than must a restricted-key algorithm. Copy protection applications require that a watermark can be read by anyone, even by potential copyright pirates, but nonetheless only the sender should be able to erase the watermark. The problem is that complete knowledge of the detection algorithm and key can imply knowledge of how to insert watermarks, and, in general a watermark can be erased by using the insertion algorithm to insert the negation of the watermark pattern. The ideal solution would be an algorithm in which knowing how to detect would not imply knowing how to insert, but this would be a true public-key algorithm, and, as pointed out above, we know of no such algorithm.

In the absence of true public watermarking, one alternative for unrestricted-key watermarking is to use an existing algorithm placed in a tamper-resistant box. However, this approach has weaknesses and other disadvantages. An attacker may be able to reverse engineer the tamper resistant box. For the consumer electronics and computer industry, the logistics of the manufacturing process are more complicated and less flexible if secret data has to be handled during design, prototyping, testing, debugging and quality control. Some of the attacks to be described in Section 18.6 exploit the fact that algorithms which are inherently "secret key" in nature, are used in an environment where public detection properties are desired, i.e. access to the key is almost completely unrestricted.

An example of restricted-key watermarking is in the broadcast industry which uses watermarks to automatically monitor and log the radio music that is broadcast. This facilitates the transfer of airplay royalties to the music industry. In a scenario where monitoring receivers are located "in the field", the watermark embedding system as well as any and all receiving monitors can be owned and operated by the royalty collection agency. However, in practice radio stations are more interested in reducing the work load of their studio operators (typically a single disk jockey) than to intentionally evade royalty payments and mostly use watermark readers themselves to create logs. As already mentioned in the introduction, watermarking of television news clips are under research, for instance in the European VIVA

project.

A similar scenario is used for a service in which images are watermarked and search robots scan the Internet to find illegally posted copies of these images. In this scenario it is not a fundamental problem that the watermark detector contains sensitive secret data, i.e., a detection key, that would reveal how the watermark can be erased. Potential attackers do not, in principle, have access to a watermark detector. However, a security threat occurs if a detector may accidentally fall into the hands of a malicious user. Moreover, the watermark solution provider may offer a service to content publishers to verify on-line whether camera-ready content is subject to copy restriction. Such an on-line service could be misused in an attack to deduce the watermark secrets.

False positive rate In most applications, it is necessary to distinguish between data that contains watermarks and data that doesn't. The false positive rate of a watermark detection system is the probability that it will identify an unwatermarked piece of data as containing a watermark. The seriousness of such an error depends on the application. In some applications, it can be catastrophic.

For example, in the copy control application considered for DVD, a device will refuse to play video from a non-factory-recorded disk if it finds a watermark saying that the data should never be copied. If a couple's wedding video (which would doubtless be unwatermarked and would not be on a factory recorded disk) is incorrectly identified as watermarked, then they will never be able to play the disk. Unless such errors are extremely rare, false positives could give DVD players a bad reputation that would seriously damage the market for them. The estimates of most companies competing to design the watermarking method used in DVD place the acceptible false positive rate at one false positive in several tens or hundreds of billions of distinct frames.

Modification and multiple watermarks In some circumstances, it is desirable to alter the watermark after insertion. For example, in the case of digital video discs, a disc may be watermarked to allow only a single copy. Once this copy has been made, it is then necessary to alter the watermark on the original disc to prohibit further copies. Changing a watermark can be accomplished by either (i) removing the first watermark and then adding a new one or (ii) inserting a second watermark such that both are readable, but one overrides the other. The first alternative does not allow a watermark to be tamper

resistant since it implies that a watermark is easily removable. Allowing multiple watermarks to co-exist is preferable and also facilitates the tracking of content from manufacturing to distribution to eventual sales, since each point in the distribution chain can insert their own unique watermark.

There is however a security problem related with multiple watermarks as explained in Section 18.6.6. If no special measures are taken the availability of a single original with different watermarks will allow a clever pirate to retrieve the unmarked original signal by statistical averaging or more sophisticated methods [7, 10].

Data payload Fundamentally, the data payload of a watermark is the amount of information it contains. As with any method of storing data, this can be expressed as a number of bits, which indicates the number of distinct watermarks that might be inserted into a signal. If the watermark carries N bits, then there are 2^N different possible watermarks. It should be noted, however, that there are actually $2^N + 1$ possible values returned by a watermark detector, since there is always the possibility that no watermark is present.

In discussing the data payload of a watermarking method, it is important to distinguish between the number of distinct watermarks that may be inserted, and the number of watermarks that may be detected by a single iteration with a given watermark detector. In many watermarking applications, each detector need not test for all the watermarks that might possibly be present. For example, several companies might want to set up web-crawlers that look for the companies' watermarks in images on the web. The number of distinct possible watermarks would have to be at least equal to the number of companies, but each crawler could test for as few as one single watermark. A watermarking system tailored for such an application might be said to have a payload of many bits, in that many different watermarks are possible, but this does not mean that all the bits are available from any given detector.

Computational cost As with any technology intended for commercial use, the computational costs of inserting and detecting watermarks are important. This is particularly true when watermarks need to be inserted or detected in real-time video or audio.

The speed requirements are highly application dependent. In general, there is often an asymmetry between the requirement for speed of insertion and speed of detection. For example, in the DIVX fingerprinting application, watermarks must be inserted in real-time by

inexpensive hardware – typically single chips costing only a few dollars each – while they may be detected, in less than real-time, by professional equipment costing tens of thousands of dollars. On the other hand, in the case of copy-control for DVD, it is the detection that must be done in real-time on inexpensive chips, while the insertion may be done on high-cost professional equipment. Note that, in cases like DVD where we can afford expensive inserters, it can actually be desirable to make the inserters expensive, since an inserter is often capable of removing a watermark, and we want them to be difficult for pirates to obtain or reproduce.

Another issue to consider in relation to computational cost is the issue of scalability. It is well known that computer speeds are approximately doubling every eighteen months, so that what looks computationally unreasonable today may very quickly become a reality. It is therefore very desirable to design a watermark whose detector and/or inserter is scalable with each generation of computers. Thus, for example, the first generation of detector might be computationally inexpensive but might not be as reliable as next generation detectors that can afford to expend more computation to deal with issues such as geometric distortions.

Standards In some application scenarios watermark technology needs to be standardized to allow global usage. An example where standardization is needed is DVD. A copy protection system based on watermarks is under consideration that will require every DVD player to check for a watermark in the same way. However, a standardized detection scheme does not necessarily mean that the watermark insertion method also needs to be standardized. This is very similar to the standardization activities of MPEG, where the syntax and the semantics of the MPEG bitstream is fixed, but not the way in which an MPEG bitstream is derived from baseband video. Thus, companies may try to develop embedding systems which are superior with respect to robustness or visibility.

18.4 Example of a Watermarking method

To evaluate watermarking properties and detector performance in more detail, we now present a basic class of watermarking methods.

Mathematically, given an original image S_0 and a watermark W, the watermarked image, S_w, is formed by $S_w = S_0 + f(S_0, W)$ such that the

watermarked image S_w is constrained to be visually identical (or very similar) to the original unwatermarked image S_0.

In theory, the function f may be arbitrary, but in practice robustness requirements pose constraints on how f can be chosen. One requirement is that watermarking has to be robust to random noise addition. Therefore many watermark designers opt for a scheme in which image S_0 will result in approximately the same watermark as a slightly altered image $S_0 + \epsilon$. In such cases $f(S_0, W) \approx f(S_0 + \epsilon, W)$

For an unrestricted-key watermark, detection of the watermark, W, is typically achieved by correlating the watermark with some function, g, of the watermarked image. Thus, the key simply is a pseudo-random number sequence, or a seed for the generator that creates such sequence, that is embedded in all images.

Example: In its basic form, in one half of the pixels the luminance is increased by one unit step while the luminance is kept constant [11] or decreased by one unit step [12] in the other half. Detection by summing luminances in the first subset and subtracting the sum of luminances in the latter subset is a special case of a correlator. One can describe this as $S_w = S_0 + W$, with $W \in R^N$, and where $f(S_0, W) = W$. The detector computes $S_w \cdot W$, where $\cdot$ denotes the scalar product of two vectors.

If W is chosen at random, then the distribution of $S_0 \cdot W$ will tend to be quite small, as the random $\pm$ terms will tend to cancel themselves out, leaving only a residual variance. However, in computing $W \cdot W$ all of the terms are positive, and will thus add up. For this reason, the product $S_w \cdot W = S_0 \cdot W + W \cdot W$ will be close to $W \cdot W$. In particular, for sufficiently large images, it will be large, even if the magnitude of S_0 is much larger than the magnitude of W. It turns out that the probability of making an incorrect detection can be expressed as the complementary error function of the square root of the ratio $W \cdot W$ over the variance in pixel luminance values. This result is very similar to expressions commonly encountered in digital transmission over noisy radio channels. Elaborate analyses of the statistical behavior of $I \cdot W$ and $W \cdot W$ are typically found in spread-spectrum oriented papers, such as [7, 8, 13, 14, 15, 16].

18.5 Robustness to signal transformations

Embedding a copy flag in ten seconds of NTSC video may not seem difficult since it only requires the embedding of 4-bits of information in a data stream. The total video data is approximately $720 \times 480 \times 30 \times 10$. This is over 100Mbytes prior to MPEG compression. However, the constraints of (i) maintaining image fidelity and (ii) survive common signal transfor-

mations, can be severe. In particular, many signal transformations cannot be modeled as a simple linear additive noise process. Instead, such processes are highly spatially correlated and may interact with the watermark in complex ways.

There are a number of common signal transformations that a watermark should survive, e.g. affine transformations, compression/re-compression, and noise. In some circumstances, it may be possible to design a watermark that is completely invariant to a particular transformation. For example, this is usually the case for translational motions. However, scale changes are often much more difficult to design for and it may be the case that a watermark algorithm is only robust to small perturbations in scale. In this case, a series of attacks may be mounted by identifying the limits of a particular watermarking scheme and subsequently finding a transformation that is outside of these limits yet maintains adequate image fidelity.

18.5.1 Affine transformations

Shifts over a few pixels can cause watermarking detectors to miss the presence of watermark. The problem can be illustrated by our example watermarking scheme. Suppose one shifts S_w by one pixel, obtaining $S_{w,s}$. Let $S_{w,s}$ and W_s denote the similarly shifted versions of S_0 and W. Then $S_{w,s} \cdot W = I_s \cdot W + W_s \cdot W$. As before, the random +/- terms in $S_{w,s} \cdot W$ will tend to cancel themselves out. However, the $W_S \cdot W$ terms will also cancel themselves out, if each +/- value was chosen independently. Hence, $S_{w,s} \cdot W$ will have small magnitude and the watermark will not be detected.

Typical analog VHS recorders cause shifting over a small portion of a line, but enough to cause a shift of several pixels or even a few DCT blocks. Recorder time jitter and tape wear randomly stretch an image. Even if the effects are not disturbing to a viewer, it may completely change the alignment of the watermark with respect to pixels and DCT block boundaries.

There are a number of defenses against such attacks. Ideally, one would like to reverse the affine transformations. Given an original, a reasonable approximation to the distortion can be computed. With unrestricted-key watermarks, and in particular the "do not copy" application, no original is available. A secondary signal, i.e. a registration pattern, may be inserted into the image whose entire purpose is to assist in reversing the transformation. However, one can base attacks on this secondary signal, removing or altering it in order to block detection of the watermark. Another alternative is to place watermark components at key visual features of the image, e.g. in patches whose average luminosity is at a local maximum. Finally, one can insert the watermark into features that are transformation invari-

ant. For example, the magnitudes of Fourier coefficients are translation invariant.

In some applications, it may be assumed that the extent of the affine transformation is minor. Particularly if the watermark predominantly resides in perceptually relevant low-frequency components, the autocorrelation $W_s \cdot W$ can be sufficiently large for sufficiently small translations. A reliability penalty associated with low-pass watermarking is derived in [13].

18.5.2 Noise addition

A common misunderstanding is that a watermark of small amplitude can be removed by adding random noise of a similar amplitude. On the contrary, correlation detectors appear very robust against addition of a random noise term ϵ. For instance if $f(I, W) = W$ one can describe the attacked image as $S_{w,s} = S_0 + \epsilon + W$. The detector computes $S_w \cdot W$. The product $S_w \cdot W = S_0 \cdot W + \epsilon \cdot W + W \cdot W$. If the watermark was designed with $W \cdot W$ largely exceeding the statistical spreading in $I \cdot$ W, it will mostly also largely exceed the statistical spreading in $\epsilon \cdot W$. In practice, noise mostly is not a serious threat unless (in the frequency components of relevance) the noise is large compared to image I or if the noise is correlated with the watermark.

18.5.3 Spatial filtering

Most linear filters for image processing create a new image by taking a linear combination of surrounding pixels. Watermark detection can be quite reliable after such filtering, particularly after edge-enhancement type of filters [14]. Such filters typically amplify the luminance of the original image and subtract shifted versions of the surroundings. In effect, redundancy in the image is cancelled and randomness of the watermark is exaggerated. One the other hand, smoothing and low-pass filtering often reduce the reliability of a correlator watermark detector.

18.5.4 Digital compression

MPEG video compression accurately transfers perceptually important components, but coarsely quantizes high image components with high frequency components. This process may severely reduce the detectability of a watermark, particularly if that resides in high spatial frequencies. Such MPEG compression is widely used in digital television and on DVD discs.

Digital recorders may not always make a bit exact copy. Digital recorders will, at least initially, not contain sophisticated signal processing facilities.

For recording of MPEG streams onto media with limited storage capacity, the recorder may have to reduce the bit rate of the content.

For video recorders that re-compress video, image quality usually degrades significantly, as quantization noise is present, typically with large high frequency components. Moreover, at high frequencies, image and watermark components may be lost. In such cases, the watermark may be lost, though the video quality may also be significantly degraded.

18.6 Tamper Resistance

In this section, we describe a series of attacks that can be mounted against a watermarking system.

18.6.1 Attacks on the content

Although several commercially available watermarking scheme are robust to many types of transformation (e.g. rotation, scaling etc), these often are not robust to combinations of basic transformations, such as scaling, cropping and a rotation. Several tools have been created by hackers that combine a small non-linear stretching with spatial filtering [17].

18.6.2 Attacks by statistical averaging

An attacker may try to estimate the watermark and subtract this from a marked image. Such an attack is particularly dangerous if the attacker can find a generic watermark, for instance one with $W = f(S_0, W)$ not depending significantly on the image S_0. Such an estimate W of the watermark can then be used to remove a watermark from any arbitrary marked image, without any further effort for each new image or frame to be "cleaned".

The attacker may separate the watermark W by adding or averaging multiple images, e.g. multiple successive marked images $S_0 + W, S_1 + W, \ldots, S_N + W$ from a video sequence. The addition of N such images results in $NW + \sum_i S_i$, which tends to NW for large N and sufficiently many and sufficiently independent images $S_0, S_1, \ldots, S_N$.

A countermeasure is to use at least two different watermarks W_1 and W_2 at random, say with probability p_1 and p_2 where $p_2 = 1 - p_1$, respectively. The above attack then only produces $p_1 W_1 + (1 - p_1) W_2$, without revealing W_1 or W_2. However a refinement of the attack is to compute weighted averages, where the weight factor is determined by a (possibly unreliable but better than random) guess of whether a particular image contains one watermark or the other. For instance, the attacker may put an image in

category $i (i \in \{1, 2\})$ if he believes that this image contains watermark W_i. Let P_ϵ denote the probability that an image is put into the wrong category. Then, after averaging a large number (N_1) of images from category 1, the result converges to $x_1 = N_1 p_1 (1 - P_\epsilon) W_1 + N_1 (1 - p_1)(P_\epsilon) W_2$. Similarly the sum of N_2 images in category 2 tends to $x_2 = N_2 p_1 P_\epsilon W_1 + N_2 (1 - p_1)(1 - P_\epsilon) W_2$. Computing the weighted difference gives

$$\frac{x_1}{N_1} - \frac{x_2}{N_2} = p_1 (1 - 2P_\epsilon) W_1 - (1 - p_1)(1 - 2P_\epsilon) W_2.$$

Hence for any $P_\epsilon \neq 1/2$, i.e., for any selection criterion better than a random one, the attacker can estimate both the sum and difference of $p_1 W_1$ and $(1 - p_1) W_2$. This reveals W_1 and W_2.

18.6.3 Exploiting the presence of a watermark detector device

For unrestricted-key watermarks, we must assume that the attacker at least has access to a "black box" watermark detector, which indicates whether the watermark is present in a given signal. Using this detector, the attacker can probably learn enough about the detection region, in a reasonable amount of time, to reliably remove the watermark.

The aim of the attack is to experimentally deduce the behavior of the detector, and to exploit this knowledge to ensure that a particular image does not trigger the detector. For example, if the watermark detector gives a soft decision, e.g. a continuous reliability indication when detecting a watermark, the attacker can learn how minor changes to the image influence the strength of the detected watermark. That is, modifying the image pixel-by-pixel, he can deduce the entire correlation function or other watermark detection rule. Interestingly, such attack can also be applied even when the detector only reveals a binary decision, i.e. present or absent. Basically the attack [18, 19] examines an image that is at the boundary where the detector changes its decision from "absent" to "present". For clarity the reader may consider a watermark detector of the correlator type; but this is not a necessary condition for the attack to work. For a correlator type of detector, our attack reveals the correlation coefficients used in the detector (or at least their sign).

For example:

1. Starting with a watermarked image, the attacker creates a test image that is near the boundary of a watermark being detectable. At this point it does not matter whether the resulting image resembles the original or not. The only criterion is that minor modifications

to the test image cause the detector to respond with "watermark" or "no watermark" with a probability that is sufficiently different from zero or one. The attacker can create the test image by modifying a watermarked image step-by-step until the detector responds "no watermark found". A variety of modifications are possible. One method is to gradually reduce the contrast in the image just enough to drop below the threshold where the detector reports the presence of the watermark. An alternative method is to replace more and more pixels in the image by neutral grey. There must be a point where the detector makes the transition from detecting a watermark to responding that the image contains no watermark. Otherwise this step would eventually result in an evenly grey colored image, and no reasonable watermark detector can claim that such image contains a watermark.

2. The attacker now increases or decreases the luminance of a particular pixel until the detector sees the watermark again. This provides the insight of whether the watermark embedder decreases or increases the luminance of that pixel.

3. This step is repeated for every pixel in the image.

4. Combining the knowledge on how sensitive the detector is to a modification of each pixel, the attacker estimates a combination of pixel values that has the largest influence on the detector for the least disturbance of the image.

5. The attacker uses the original marked image and subtracts (λ times) the estimate, such that the detector reports that no watermark is present. λ is found experimentally, such that λ is as small as possible. Moreover, the attacker may also exploit a perceptual model to minimize the visual effect of his modifications to the image.

The computational effort needed to find the watermark is much less than commonly believed. If an image contains N pixels, conventional wisdom is that an attack that searches the watermark requires an exponential number of attempts of order $O(2^N)$. A brute force exhaustive search checking all combinations with positive and negative sign of the watermark in each pixel results in precisely 2^N attempts. The above method shows that many watermarking methods can be broken much faster, namely in $O(N)$, provided a device is available that outputs a binary (present or absent) decision as to the presence of the watermark.

We can, however, estimate the computation required to learn about the detection region when a black box detector is present, and this opens up

the possibility of designing a watermarking method that specifically makes the task impractical.

Linnartz [19] has suggested that a probabilistic detector[‡] would be much less useful to an attacker than a deterministic one. If properly designed, a probabilistic detector would teach an attacker so little in each iteration that the task would become impractical.

A variation of the attack above which also works in the case of probabilistic detectors is presented in [20] and [21]. Similar to the attack above the process starts with the construction of a signal S_θ at threshold of detection. The attacker than chooses a random perturbation V and records the decision of the watermark detector for $S_\theta + V$. If the detector sees the watermark, the perturbation V is considered an estimation of the watermark W. If the detector does not see the watermark the negation $-V$ is considered an estimation of the watermark. By repeating this perturbation process a large number of times and summing all intermediate estimates a good approximation of the watermark W can be obtained. It can be shown that the accuracy of the estimation is $\mathcal{O}(\sqrt{\frac{J}{N}})$ where J is the number of trials and N is the number of samples. In particular it follows that for a fixed accuracy κ the number of trials J is linear with the number of samples N. A more detailed analysis also shows that the number of trials is proportional to the square of the width of the threshold zone (i.e. the zone where the detector takes probabilistic decisions). The designer of a probabilistic watermark detector therefore faces the trade-off between a large threshold zone (i.e. a high security), a small false negative rate (i.e. a small upper bound of the threshold zone) and a small false positive rate (i.e. a large lower bound of the threshold zone).

18.6.4 Attacks based on the presence of a watermark inserter

If the attacker has access to a watermark inserter, this provides further opportunities to break the security. Attacks of this kind are relevant to copy control in which copy generation management is required, i.e. the user is permitted to make a copy from the original source disc but is not permitted to make a copy of the copied material - only one generation of copying is allowed. The recorder should change the watermark status from

[‡]A probabilistic detector is one in which two thresholds exist. If the detector output is below the lower threshold then no watermark is detected. Similarly, if the detector output is above the higher threshold then a watermark is detected. However, if the detector output lies between the two thesholds, then then the decision as to whether the watermark is present or absent is random.

"one-copy allowed" to "no more copies allowed". The attacker has access to the content before and after this marking. That is, he can create a difference image, by subtracting the unmarked original from the marked content. This difference image is equal to $f(S_0, W)$. An obvious attack is to pre-distort the original to undo the mark addition in the embedder. That is, the attacker computes $I - f(S_0, W)$ and hopes that after embedding of the watermark, the recorder stores

$$S_0 - f(S_0, W) + f(S_0 - f(S_0, W), W)$$

which is likely to approximate S_0. The reason why most watermarking methods are vulnerable to this attack is that watermarking has to be robust to random noise addition. If, for reasons discussed before,

$$f(S_0, W) \approx f(S_0 + \epsilon, W),$$

and because watermarks are small modifications themselves, $f(S_0, W) \approx f(S_0 - f(S_0, W), W)$. This property enables the above pre-distortion attack.

18.6.5 Attacks on the copy protection system

The forgoing discussion of tamper-resistance has concentrated only on the problem of removing a watermark from a given signal. We have not discussed ways of circumventing systems that are based on watermarking. In many applications, it is far easier to thwart the purpose of the watermark than it is to remove the watermark. For example, Craver *et al* [1], discuss ways in which watermarks that are used to identify media ownership might be thwarted by inserting conflicting watermarks into the signal so as to make it impossible to determine which watermark identifies the true owner. Cox and Linnartz [18, 22] discuss several methods of circumventing watermarks used for copy control.

The most trivial attack is to tamper with the output of the watermark detector and modify it in such a way that the copy control mechanism always sees a "no watermark" detection, even if a watermark is present in the content. Since hackers and pirates more easily can modify (their own!) recorders but not their customers' players, playback control is a mechanism that detects watermarks during the playback of discs. The resulting tape or disc can be recognized as an illegal copy if playback control is used.

Copy protection based on watermarking content has a further fundamental weakness. The watermark detection process is designed to detect the watermark when the video is perceptually meaningful. Thus, a user may apply a weak form of scrambling to copy protected video, e.g. inverting the pixel intensities, prior to recording. The scrambled video is unwatchable

and the recorder will fail to detect a watermark and consequently allow a copy to be made. Of course, on playback, the video signal will be scrambled, but the user may then simple invert or descramble the video in order to watch a perfect and illegal copy of a video. Simple scrambling and descrambling hardware would be very inexpensive and manufacturers might argue that the devices serve a legitimate purpose in protecting a user's personal video. Similarly, digital MPEG can easily be converted into a file of seemingly random bits. One way to avoid such circumvention for digital recording is to only allow the recording of content in a recognized file format. Of course this would severely limit the functionality of the storage device.

18.6.6 Collusion attacks

If the attacker has access to several versions of the signal, $S_{w_1}, S_{w_2} \ldots S_{w_N}$, each with a different watermark, but each perceptually equivalent to S_0, then he/she can learn much more about the region of signals that are equivalent to S_0, since it will be well approximated by the intersection of the regions of signals that are equivalent to the watermarked signals. This gives rise to "collusion attacks", in which several watermarked signals are combined to construct an unwatermarked signal. The attacker's knowledge of the detection region is under our direct control. In the case of a restricted-key watermark, she/he has no knowledge of this region at all. This makes it extremely difficult to tamper with restricted-key watermarks. The best an attacker can do is to find a signal that is as far from the watermarked signal as possible, while still likely to be within the range of signals perceptually equivalent to S_0, and to hope that this distant signal is outside the detection range. In the case of a collusion attack, this job is made easier, because the hacker can use the multiple watermarked versions of the signal to obtain closer and closer approximations to S_0, which definitely is not watermarked. However, whether or not the attacker has the advantage of making a collusion attack, he/she can never be sure whether the attack succeeded, since the information required to test for the watermark's presence is not available. This should help make security systems based on restricted-key watermarks more effective. Resistance to collusion attacks is also a function of the structure of the watermark, as discussed in [10].

In the next section, we summarize early work on watermarking and then describe more recent work which attempts to insert a watermark into the perceptually significant regions of an image.

18.7 Proposed methods

In this section, we provide a review of watermarking methods that have been proposed. This is unlikely to be a complete list and omissions should not be interpreted as being inferior to those described here. Recent collections of papers can be found in [23, 24].

Early work on watermarking focused on hiding information within a signal but without considering the issues discussed earlier. In an application in which a covert channel between two parties is desired, tamper resistance may not be an issue if only the communicating parties are aware of the channel. Thus, early work can be thought of as steganography [25].

Turner [6] proposed inserting an identification code into the least significant bits of randomly selected words on compact discs. Decoding is accomplished by comparison with the original unwatermarked content. Although the method is straightforward, it is unlikely to be robust or tamper resistant. For example, randomizing the least significant bits of all words would remove the watermark. Oomen *at al.*[26] refined the method exploiting results from the theory of perceptual masking, dithering and noise shaping. Later van Schyndel *et al* [27] proposed a similar method as well as a spread spectrum method that linearly adds a watermark to an image.

Brassil *et al* [28] describe several methods for watermarking text, based on slightly altering the character or line spacings on a page or by adding/deleting serifs from characters. This approach is further refined in [29]. Unfortunately, as the authors note, these approaches are not resistant to tampering. For example, a malicious attacker could randomize the line or character spacing, thereby destroying the watermark. In general, text is particularly difficult to watermark based on adding noise, since optical character technology is, in principle, capable of eliminating it. An alternative approach is to insert the watermark at the symbolic level, by, for example, inserting spelling errors or by replacing words or phrases with alternatives in a predetermined manner, e.g. substituting "which" for "that". However, these approaches also appear susceptible to tampering.

Caronni [30] describes a procedure in which faint geometric patterns are added to an image. The watermark is therefore independent of the image, but because the watermark is graphical in nature, it has a spatial frequency distribution that contains perceptually significant components. However, it is unclear whether such a method is preferable to adding a pre-filtered PN noise sequence.

Tanaka *et al* [31] proposed a method to embed a signal in an image when the image is represented by dithering. Later, Matsui and Tanaka [32] suggested several different methods to encode a watermark, based on whether the image was represented by predictive coding, dithering (monotone printing) or run-lengths (fax). A DCT-based method is also proposed for video sequences. These methods make explicit use of the representation and it is unclear whether such approaches are robust or tamper resistant.

Koch *et al* [33, 34] describe several procedures for watermarking an image based on modifying pairs or triplets of frequency coefficients computed as part of the JPEG compression procedure. The rank ordering of these frequency coefficients is used to represent the binary digits. The authors select mid-range frequencies which typically survive JPEG compression. To avoid creating artifacts, the DC coefficient is not altered. Several similar methods has recently been proposed. Bors and Pitas [35] suggest an alternative linear constraint among selected DCT coefficients, but it is unclear whether this new constraint is superior to that of [33, 34]. Hsu and Wu [36] describe a method in which the watermark is a sequence of binary digits that are inserted into the mid-band frequencies of the 8×8 DCT coefficients.

Swanson *et al* [37] describe linearly adding a PN sequence that is first shaped to approximate the characteristics of the human visual system to the DCT coefficients of 8×8 blocks. In the latter two cases, the decoder requires access to the original image. It is interesting to note that a recently issued patent [38] appears to patent the general principle of extracting a watermark based on comparison of the watermarked and unwatermarked image.

Rhoads [39] describes a method in which N pseudo random (PN) patterns, each pattern having the same dimensions as the image, are added to an image in order to encode an N-bit word. The watermark is extracted by first subtracting a copy of the unwatermarked image and correlating with each of the N know PN sequences. The need for the original image at the decoder was later relaxed. While Rhoads did not explicitly recognize the important of perceptual modeling, experiments with image compression led him to propose that the PN sequences be spectrally filtered, prior to insertion, such that the filtered noise sequence was within the passband of common image compression algorithms such as JPEG.

Bender *et al* [40] describe several possible watermarking methods. In particular, "Patchwork" encodes a watermark by modifying a statistical

property of the image. The authors note that the difference between any pair of randomly chosen pixels is Gaussian distributed with a mean of zero. This mean can be shifted by selecting pairs of points and incrementing the intensity of one of the points while decrementing the intensity of the other. The resulting watermark spectrum is predominantly high frequency. However, the authors recognize the importance of placing the watermark in perceptually significant regions and consequently modify the approach so that pixel patches rather than individual pixels are modified, thereby shaping the watermark noise to significant regions of the human visual system. While the exposition is quite different from Rhoads [39], the two techniques are very similar and it can be shown that the Patchwork decoder is effectively computing the correlation between the image and a binary noise pattern, as covered in our example detector in Section 18.4.

Paatelma and Borland [41] propose a procedure in which commonly occurring patterns in images are located and target pixels in the vicinity of these patterns are modified. Specifically, a pixel is identified as a target if it is preceded by a preset number of pixels along a row that are all different from their immediate neighbors. The target pixel is then set to the value of the pixel a fixed offset away, provided the intensity difference between the two pixels does not exceed a threshold. Although the procedure appears somewhat convoluted, the condition on target pixels assures that the watermark is placed in regions that have high frequency information. Although the procedure does not explicitly discuss perceptual issues, a commercial implementation of this process is claimed to have survived through the printing process.

Holt *et al* [42] describe a watermarking procedure in which the watermark is first nonlinearly combined with an audio signal to spectrally shape it and the resulting signal is then high pass filtered prior to insertion into the original audio signal. Because of the high pass filtering, the method is unlikely to be robust to common signal distortions. However, Preuss *et al* [43] describe an improved procedure that inserts the shaped watermark into the perceptually significant regions of the audio spectrum. The embedded signaling procedure maps an alphabet of signals to a set of binary PN sequences whose temporal frequency response is approximately white. The audio signal is analyzed through a window and the audio spectrum in this window is calculated. The watermark and audio signals are then combined nonlinearly by multiplying the two spectra together. This combined signal will have a shape that is very similar to the original audio spectrum. The resulting signal is then inverse transformed and linearly

weighted and added to the original audio signal. This is referred to as spectral shaping. To decode the watermark, the decoder first applies a spectral equalizer that whitens the received audio signal prior to filtering through a bank of matched filters, each one tuned to a particular symbol in the alphabet. While the patent does not describe experimental results, we believe that this is a very sophisticated watermarking procedure that should be capable of surviving many signal distortions.

Cox *et al* [7, 8] describe a somewhat similar system for images in which the perceptually most significant DCT coefficients are modified in a non-linear fashion that effectively shapes the watermark spectrum to that of the underlying image. The decoder requires knowledge of the original unwatermarked image in order to invert the process and extract the watermark. This constraint has been subsequently relaxed. The authors also note that binary watermarks are less resistant to tampering by collusion than watermarks that are based on real valued, continuous pseudo random noise sequences.

Podilchuk and Zeng [44] describe improvements to Cox *et al* by using a more advanced perceptual model and a block based method that is therefore more spatially adapative.

O Ruanaidh [45] describe an approach similar to [7, 8] in which the phase of the DFT is modified. The authors note that phase information is perceptually more significant than the magnitude of Fourier coefficients and therefore argue that such an approach should be more robust to tampering as well as to changes in image contrast. The inserted watermark is independent of the image and is recovered using traditional correlation without the use of the original image.

Several authors [43, 33, 34, 7, 8, 13, 14], draw upon work in spread spectrum communications. Smith and Comiskey [15] analyze watermarking from a communications perspective. They propose a spread spectrum based technique that "predistorts" the watermark prior to inserting. However, the embedded signal is not a function of the image, but rather is pre-filtered based on expected compression algorithms such as JPEG. Linnartz *et al.* [13, 14], review models commonly used for detection of spread spectrum radio signals and discuss their suitability in evaluating watermark detector performance. In contrast to typical radio systems in which the signal waveform (e.g. whether it is spread or not) does not affect error performance according to the

most commonly accepted channel model,[§] the watermark detector tends to be sensitive to the spectral shape of the watermark signal. A signal-to-noise penalty is derived for placing the watermark in visually important regions, in stead of using a spectrally flat (unfiltered) PN-sequence.

18.8 Summary

We have described a basic framework in which to discuss the principals of watermarking, and outlined several characteristics of watermarks that might be desirable for various applications. We covered intentional and unintentional attacks which a watermark system may face. While a watermark may survive many signal transformations that occur in commonly used signal processing operations, resistance to intentional tampering usually is more difficult to achieve. Finally, we surveyed many of the numerous recent proposal for watermarking and attempted to identify their strengths and weaknesses.

[§]The linear time-invariant channel with additive white Gaussian noise.

Bibliography

[1] S. Craver, N. Memon, B.-L. Yeo, and M. Yeung, "Resolving rightful ownerships with invisible watermarking techniques: Limitations, attacks and implications," *IEEE Trans. on Selected Areas of Communications*, vol. 16, no. 4, pp. 573–586, 1998.

[2] R. J. Anderson and F. A. P. Petitcolas, "On the limits of steganography," *IEEE Trans. on Selected Areas of Communications*, vol. 16, no. 4, pp. 474–481, 1998.

[3] G. Simmons, "The prisoner's problem and the subliminal channel," in *Proceedings CRYPTO'83*, Advances in Cryptology, pp. 51–67, Plenum Press, 84.

[4] R. L. Rivest, "Chaffing and winnowing: Confidentiality without encryption." http:theory.lcs.mit.edu/ rivest/chaffing.txt, 1998.

[5] N. Jayant, J. Johnston, and R. Safranek, "Signal compression based on models of human perception," *Proc IEEE*, vol. 81, no. 10, 1993.

[6] L. F. Turner, "Digital data security system." Patent IPN WO 89/08915, 1989.

[7] I. Cox, J. Kilian, F. T. Leighton, and T. Shamoon, "Secure spread spectrum watermarking for images, audio and video," in *IEEE Int. Conference on Image Processing*, vol. 3, pp. 243–246, 1996.

[8] I. Cox, J. Kilian, F. T. Leighton, and T. Shamoon, "A secure, robust watermark for multimedia," in *Information Hiding: First Int. Workshop Proc.* (R. Anderson, ed.), vol. 1174 of *Lecture Notes in Computer Science*, pp. 185–206, Springer-Verlag, 1996.

[9] I. Cox and M. L. Miller, "A review of watermarking and the importance of perceptual modeling," in *Proceedings of SPIE, Human Vision & Electronic Imaging II*, vol. 3016, pp. 92–99, 1997.

[10] I. Cox, J. Kilian, F. T. Leighton, and T. Shamoon, "Secure spread spectrum watermarking for images, audio and video," *IEEE Trans. on Image Processing*, vol. 6, no. 12, pp. 1673–1687, 1997.

[11] I. Pitas and T. Kaskalis, "Signature casting on digital images," in *Proceedings IEEE Workshop on Nonlinear Signal and Image Processing*, (Neos Marmaras), June 1995.

[12] W. Bender, D. Gruhl, and N. Morimoto, "Techniques for data hiding," in *Proc. of SPIE*, vol. 2420, p. 40, February 1995.

[13] J. Linnartz, A. Kalker, and G. Depovere, "Modelling the false-alarm and missed detection rate for electronic watermarks," in *Workshop on Information Hiding, Portland, OR*, 15-17 April, 1998.

[14] G. Depovere, T. Kalker, and J.-P. Linnartz, "Improved watermark detection using filtering before correlation," in *Proceedings of the ICIP*, (Chicago), Oct. 1998. Submitted.

[15] J. R. Smith and B. O. Comiskey, "Modulation and information hiding in images," in *Information Hiding: First Int. Workshop Proc.* (R. Anderson, ed.), vol. 1174 of *Lecture Notes in Computer Science*, pp. 207–226, Springer-Verlag, 1996.

[16] J. J. Hernandez, F. Perez-Gonzalez, J. M. Rodriguez, and G. Nieto, "Performance analysis of a 2-D multipulse amplitude modulation scheme for data hiding and watermarking still images," *IEEE Trans. on Selected Areas of Communications*, vol. 16, no. 4, pp. 510–524, 1998.

[17] F. Petitcolas, R. Anderson, and M. Kuhn, "Attacks on copyright marking systems," in *Workshop on Information Hiding, Portland, OR*, 15-17 April, 1998.

[18] I. Cox and J.-P. Linnartz, "Public watermarks and resistance to tampering," in *Proceedings of the IEEE International Conference on Image Processing, CDRom*, 1997.

[19] J. Linnartz and M. van Dijk, "Analysis of the sensitivity attack against electronic watermarks in images," in *Workshop on Information Hiding, Portland, OR*, 15-17 April, 1998.

[20] T. Kalker, J. Linnartz, and M. van Dijk, "Watermark estimation through detector analysis," in *Proceedings of the ICIP*, (Chicago), Oct. 1998. Accepted.

[21] T. Kalker, "Watermark estimation through detector observations," in *Proceedings of the IEEE Benelux Signal Processing Symposium*, (Leuven, Belgium), pp. 119–122, Mar. 1998.

[22] I. J. Cox and J.-P. Linnartz, "Some general methods for tampering with watermarks," *IEEE Trans. on Selected Areas of Communications*, vol. 16, no. 4, pp. 587–593, 1998.

[23] R. Anderson, ed., *Information Hiding*, vol. 1174 of *Lecture Notes in Computer Science*, Springer-Verlag, 1996.

[24] *IEEE Int. Conf. on Image Procesing*, 1996.

[25] D. Kahn, "The history of steganography," in *Information Hiding* (R. Anderson, ed.), vol. 1174 of *Lecture Notes in Computer Science*, pp. 1–5, Springer-Verlag, 1996.

[26] A. Oomen, M. Groenewegen, R. van der Waal, and R. Veldhuis, "A variable-bit-rate buried-data channel for compact disc," in *Proc. 96th AES Convention*, 1994.

[27] R. G. van Schyndel, A. Z. Tirkel, and C. F. Osborne, "A digital watermark," in *Int. Conf. on Image Processing*, vol. 2, pp. 86–90, IEEE, 1994.

[28] J. Brassil, S. Low, N. Maxemchuk, and L. O'Gorman, "Electronic marking and identification techniques to discourage document copying," in *Proc. of Infocom'94*, pp. 1278–1287, 1994.

[29] J. Brassil and L. O'Gorman, "Watermarking document images with bounding box expansion," in *Information Hiding* (R. Anderson, ed.), vol. 1174 of *Lecture Notes in Computer Science*, pp. 227–235, Springer-Verlag, 1996.

[30] G. Caronni, "Assuring ownership rights for digital images," in *Proc. Reliable IT Systems, VIS'95*, Vieweg Publishing Company, 1995.

[31] K. Tanaka, Y. Nakamura, and K. Matsui, "Embedding secret information into a dithered multi-level image," in *Proc, 1990 IEEE Military Communications Conference*, pp. 216–220, 1990.

[32] K. Matsui and K. Tanaka, "Video-steganography," in *IMA Intellectual Property Project Proceedings*, vol. 1, pp. 187–206, 1994.

[33] E. Koch, J. Rindfrey, and J. Zhao, "Copyright protection for multimedia data," in *Proc. of the Int. Conf. on Digital Media and Electronic Publishing*, 1994.

[34] E. Koch and Z. Zhao, "Towards robust and hidden image copyright labeling," in *Proceedings of 1995 IEEE Workshop on Nonlinear Signal and Image Processing*, June 1995.

[35] A. G. Bors and I. Pitas, "Image watermarking using dct domain constraints," in *IEEE Int. Conf. on Image Processing*, 1996.

[36] C.-T. Hsu and J.-L. Wu, "Hidden signatures in images," in *IEEE Int. Conf. on Image Processing*, 1996.

[37] M. D. Swanson, B. Zhu, and A. H. Tewfik, "Transparent robust image watermarking," in *IEEE Int. Conf. on Image Processing*, 1996.

[38] D. C. Morris, "Encoding of digital information." European Patent EP 0 690 595 A1, 1996.

[39] G. B. Rhoads, "Indentification/authentication coding method and apparatus," *World Intellectual Property Organization*, vol. IPO WO 95/14289, 1995.

[40] W. Bender, D. Gruhl, N. Morimoto, and A. Lu, "Techniques for data hiding," *IBM Systems Journal*, vol. 35, no. 3/4, pp. 313–336, 1996.

[41] O. Paatelma and R. H. Borland, "Method and apparatus for manipulating digital data works." WIPO Patent WO 95/20291, 1995.

[42] L. Holt, B. G. Maufe, and A. Wiener, "Encoded marking of a recording signal." UK Patent GB 2196167A, 1988.

[43] R. D. Preuss, S. E. Roukos, A. W. F. Huggins, H. Gish, M. A. Bergamo, P. M. Peterson, and D. A. G, "Embedded signalling." US Patent 5,319,735, 1994.

[44] C. I. Podilchuk and W. Zeng, "Image-adaptive watermarking using visual models," *IEEE Trans. on Selected Areas of Communications*, vol. 16, no. 4, pp. 525–539, 1998.

[45] J. J. K. O. Ruanaidh, W. J. Dowling, and F. Boland, "Phase watermarking of digital images," in *IEEE Int. Conf. on Image Processing*, 1996.